Reprint Series
5 March 1993, Volume 259, pp. 1430–1433

SCIENCE

On the Application of the Minimal Principle to Solve Unknown Structures

Russ Miller,* George T. DeTitta, Rob Jones David A. Langs,
Charles M. Weeks, and Herbert A. Hauptman

On the Application of the Minimal Principle to Solve Unknown Structures

Russ Miller,* George T. DeTitta, Rob Jones, David A. Langs, Charles M. Weeks, Herbert A. Hauptman

The Shake-and-Bake method of structure determination is a new direct methods phasing algorithm based on a minimum-variance, phase invariant residual, which is referred to as the minimal principle. Previously, the algorithm had been applied only to known structures. This algorithm has now been applied to two previously unknown structures that contain 105 and 110 non-hydrogen atoms, respectively. This report focuses on (i) algorithmic and parametric optimizations of Shake-and-Bake and (ii) the determination of two previously unknown structures. Traditional tangent formula phasing techniques were unable to unravel these two new structures.

The Shake-and-Bake procedure (*1*, *2*) has been designed and implemented on a variety of computing platforms for the purpose of determining crystal structures by means of minimizing a recently proposed minimal function (*2*–*5*). The focus of this report is on: (i) algorithmic and parametric optimizations we have made to the basic algorithm (*1*), based on successful applications of Shake-and-Bake to 14 known structures over six space groups, ranging from 25 to 127 non-hydrogen atoms in the asymmetric unit cell, and (ii) the application of this modified algorithm to solve two previously unknown structures. In particular, the algorithmic and parametric optimizations we present were guided predominantly by the experimentation on three known structures, namely, the 28 non-hydrogen atom 9α-methoxycortisol (*6*), the 84 non-hydrogen atom isoleucinomycin (*7*), and the 127 non-hydrogen atom isoleucinomycin analog (*8*).

The two previously unknown structures solved by our algorithm are two polymorphic forms of a cycloheptapeptide, ternatin(I), a 110 non-hydrogen atom structure (*9*), and ternatin(II), a 105 non-hydrogen atom structure (*10*).

The minimal function has not been applied previously to an unknown complex structure. A considerable effort to solve the ternatin(I) structure by traditional tangent formula methods proved unsuccessful. Approaches taken included the testing of ~50,000 randomly generated phase sets (*11*) as well as an additional 500,000 permuted phase sets. Molecular replacement (*12*) was also invoked, but the models considered were constructed with seven L-amino acids, and as it turns out, the structure was later found to contain three L- and four D-configuration amino acids. [Details of the two new structures are given in (*13*).] It remains to be demonstrated whether these two structures could have been as easily determined by several other new promising techniques that are currently being developed, including the Sayre equation tangent formula (*14*), phase annealing (*15*), and low-density elimination (*16*). In any event, the Shake-and-Bake algorithm was able to determine each structure in ~70 min of CPU time on a Connection Machine CM-5.

A general introduction to the crystallographic phase problem is given in Box 1. We have recently proposed that a particularly simple function of the phases takes on its constrained minimal value for the correct set of phases. A brief review of the minimal principle is given in Box 2. The minimal principle states that $R(P) < R(S)$ for N atom structures $S \neq P$ (the given structure). In other words, among all phases ϕ that satisfy the necessary identities, those corresponding to the true structure P, minimize $R(\phi)$. Inspection of the minimal function R shows it to be a weighted sum of squares of residuals, that is, the differences between cosines and their expected values, the ratios of the Bessel functions I_1/I_0. The known conditional probability distributions of the triple $T_{\mathbf{HK}}$, given the three magnitudes $|E|$ of Eq. 4, and the known conditional distributions of the quartet $Q_{\mathbf{LMN}}$ of Eq. 5, lead directly to the expected values of the corresponding cosines, $I_1(A_{\mathbf{HK}})/I_0(A_{\mathbf{HK}})$ and $I_1(B_{\mathbf{LMN}})/I_0(B_{\mathbf{LMN}})$, respectively. In addition, these same distributions show that the reciprocals of their variances are strongly correlated with $A_{\mathbf{HK}}$ and $|B_{\mathbf{LMN}}|$, respectively. Thus, in analogy with the principle of least squares, the minimal principle, which attempts to minimize the weighted sum of squares of residuals, Eq. 1, becomes plausible. Furthermore, it can be shown rigorously that the values of $R(\phi)$, when the phases are set equal to their true values for any choice of origin and enantiomorph, are indeed smaller than the values of $R(\phi)$ when the phases ϕ are chosen "at random."

Although a number of standard minimization techniques exist, including simulated annealing (*17*) and genetic algorithms (*18*), such techniques are targeted at minimization with respect to the range of a

R. Miller, G. T. DeTitta, D. A. Langs, C. M. Weeks, H. A. Hauptman, The Medical Foundation of Buffalo, 73 High Street, Buffalo, NY 14203.
R. Jones, Thinking Machines Corporation, 245 First Street, Cambridge, MA 02142.

*To whom correspondence should be addressed. Currently on sabbatical. Permanent address: Department of Computer Science, State University of New York at Buffalo, Buffalo, NY 14260.

function. Therefore, it does not appear that such techniques can be applied to our function to produce a solution. However, we have recently developed a computationally intensive solution strategy, which we call Shake-and-Bake, targeted at minimization with respect to the range of our function while maintaining the integrity of the parameters with respect to the domain of the function.

Notice that for a structure consisting of M independent atoms in the asymmetric unit, we must determine the 3M variables that define the fractional atomic positions in the asymmetric unit of the crystallographic unit cell. To do so, we might need to secure values for ~10M phases, which in turn might occur in some $O(M^2)$ triples and $O(M^3)$ quartets. Recasting the minimal principle from a function of invariants $R(\Phi)$ to one of structures $R(S)$ greatly simplifies our search. That is, we are in the "enviable" position of having to search for the minimum of a function of only some hundreds or thousands of variables. In order to proceed in this direction, we need a way to impose the constraints implicit in the reduction of the problem from $R(\Phi)$ to $R(S)$.

We can now consider a likely form of the constraints on the phases. We know that permissible (feasible) solutions to the phase problem (that is, sensible phase sets) should yield physically reasonable electron density maps. In particular, those maps should be everywhere nonnegative and should contain local areas of high electron density associated with atomic positions. We realized that we could impose these twin constraints on the phases by the process of Fourier inversion. That is, the minimization of $R(\Phi)$ would be allowed to proceed to some limited degree with the refined, but unconstrained, phases used (together with observed amplitudes $|E|$) to calculate a Fourier map. The resulting "structure" (in reality, the positions of the M largest nonnegative density features in the map) would in turn be used to calculate structure factor amplitudes and constrained phases. The phases are constrained in the

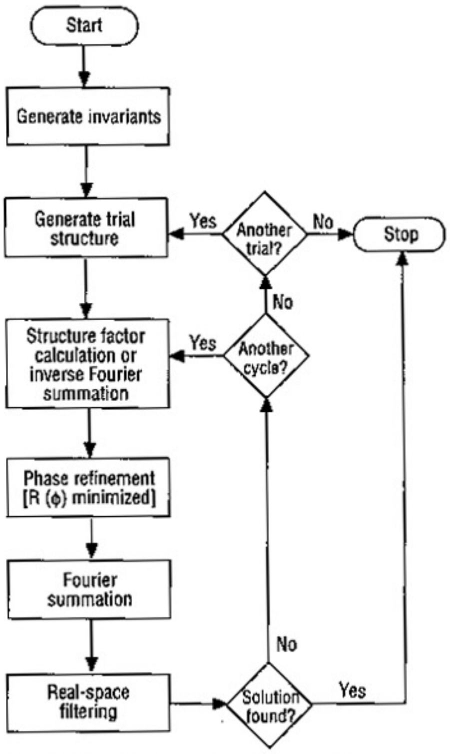

Fig. 1. The Shake-and-Bake algorithm.

```
comment: Let $\langle\phi\rangle$ be the entire set of phases
comment: Let $\phi_j$ be the $j^{th}$ phase in $\langle\phi\rangle$
Shift = Initial_Phase_Shift
For $i = 1$ to Number_of_Passes
    For $j = 1$ to Number_of_Phases
        Phase = $\phi_j$
        $R_1 = R(\langle\phi\rangle)$
        $\phi_j$ = Phase + Shift
        $R_2 = R(\langle\phi\rangle)$
        $\phi_j$ = Phase − Shift
        $R_3 = R(\langle\phi\rangle)$
        if $\min(R_1, R_2, R_3) = R_1$ then $\phi_j$ = Phase
        if $\min(R_1, R_2, R_3) = R_2$ then $\phi_j$ = Phase + Shift
        if $\min(R_1, R_2, R_3) = R_3$ then $\phi_j$ = Phase − Shift
    end{$j$-loop}
    Shift = Shift/2
end{$i$-loop}
```

Fig. 2. Pseudo-code of the global binary search routine used for performing the local minimization.

Box 1. The crystallographic phase problem. The single-crystal x-ray diffraction technique of structure determination is aimed at providing a three-dimensional map of the positions of atoms in a crystal, thereby securing unambiguous information about the architecture of a given molecule. The three stages of an x-ray diffraction experiment are:

1) The growth of suitable single crystals of the substance to be studied;
2) The measurement of x-ray diffraction data; and
3) Unraveling the molecular structure so that it agrees with the diffraction data.

The last step is frequently computationally intensive and is the focus of this research.

In the experiment (step 2, above), the crystal is oriented with respect to the x-ray beam, so that an individual diffracting plane is brought into the Bragg condition and the intensity of the diffracted photons is recorded. This process is repeated anywhere from a few hundred to a few million times, depending on the size of the structure to be determined, as individual diffracting planes are brought into the Bragg condition. Each scattered beam, called a reflection, is characterized by a location on a three-dimensional grid, or reciprocal lattice, corresponding to the orientation of the crystal and the angle which the diffracting plane makes with the incoming x-ray beam. Because the grid constitutes a true lattice, each reflection can be labeled by three integers, the Miller indices, which denote the location of the reflection on the reciprocal lattice relative to a common origin. The intensity of each reflection is related to the efficiency with which a Bragg plane diffracts x-rays. The intensity of an individual reflection is related to the density of electrons in the near vicinity of the Bragg plane. The underlying atomic arrangement in a crystal is related to the intensities and locations of the Bragg reflections by a three-dimensional Fourier transformation. We use the term real space to refer to the atomic arrangement of the crystal and the term reciprocal space to refer to the intensities and locations of the reflections.

It would seem that all of the information necessary to unravel the structure of molecules in crystals is assembled once the diffraction experiment is concluded. Unfortunately, the data produced from this experiment do not provide all of the information necessary to complete the structure. The three-dimensional atomic coordinates of the crystal are calculated by a three-dimensional Fourier transform in which the amplitudes, positions, and phases of the reflections are used. The experiment yields the amplitudes and positions of the Fourier components, but not their phases. It is the determination of these missing phases that constitutes the phase problem of x-ray crystallography.

Early analyses of the phase problem led many to believe that the problem was in principle unsolvable. An infinity of Fourier transformation maps could be had that fit the experimental results; they would differ only in the set of phases used to reconstruct the atomic arrangement. On the other hand, because a small number of structural arrangements had been ascertained by a trial and error method, it seemed that there must be a solution to the phase problem.

Two physical constraints make the problem not only solvable, but in principle greatly overdetermined. One is the hard constraint that for a Fourier transformation to be physically meaningful it must lead to a map in which the calculated electron density is everywhere nonnegative. The other is a softer constraint that the electron density about atoms in molecules (whether in crystals or in the gas phase) is strongly concentrated about the atomic centers (the nuclei). "Nonnegativity" and "atomicity" were two important principles in the earliest formulations of direct methods. In a direct-methods attack on the phase problem, probabilistic theories are used to relate the phases, or more precisely certain linear relations among the phases, which are called structure invariants, to the measured intensity data.

sense that they map to a trial structure in the known space group, with atomicity and nonnegativity explicitly imposed, and that the peaks of the map correspond to the known number of atoms M to be located. Thus, the needed constraints would not actually hold in the minimization procedure itself, but would be used to adjust the refined phases to values that do obey the constraints.

Our Shake-and-Bake solution strategy allows a simple, local minimization technique to be applied in reciprocal space, while indirectly applying the aforementioned constraints in real space (Fig. 1). In this manner, we hope to produce solutions by creating an arbitrary, yet chemically sound, structure and allow it to gradually migrate toward the correct structure by local perturbations that result in increasingly smaller values of the minimal function.

Although we continue to explore a variety of minimization techniques, including gradient descent and parameter shift, the local minimization technique used in the application of Shake-and-Bake to the ternatin structures is a global binary search routine (19). This decision was based in part on experimental evidence with respect to the 84-atom isoleucinomycin structure, which shows that the function R is monotonic, or at worst bitonic, with respect to an individual phase. This binary search routine visits each of the phases in sequence a fixed number of times. During each visit, the current value of a phase, as well as that value adjusted by a predetermined amount in both the positive and negative directions, are considered with respect to the minimal function. The best of these three values (that is, the value that produces the smallest value of the minimal function) is chosen as the (potentially) new value of the phase. An overview of this routine is given in Fig. 2, where for our application, the initial phase shift was set to 90°, and the number of passes made through the entire set of phases was five.

Box 2. The minimal principle. We assume a crystal structure P to be fixed, but unknown a priori. The normalized structure factor magnitudes $|E|$ are also assumed to be known. The function to be minimized, the so-called minimal function, is defined initially as a function, $R(\Phi)$, of the structure invariants $T_{\mathbf{HK}}$ and $Q_{\mathbf{LMN}}$.

$$R(\Phi) = \frac{\sum\limits_{\mathbf{H,K}} A_{\mathbf{HK}}\left\{\cos T_{\mathbf{HK}} - \frac{I_1(A_{\mathbf{HK}})}{I_0(A_{\mathbf{HK}})}\right\}^2 + \sum\limits_{\mathbf{L,M,N}} |B_{\mathbf{LMN}}|\left\{\cos Q_{\mathbf{LMN}} - \frac{I_1(B_{\mathbf{LMN}})}{I_0(B_{\mathbf{LMN}})}\right\}^2}{\sum\limits_{\mathbf{H,K}} A_{\mathbf{HK}} + \sum\limits_{\mathbf{L,M,N}} |B_{\mathbf{LMN}}|} \quad (1)$$

We define

$$T_{\mathbf{HK}} = \phi_{\mathbf{H}} + \phi_{\mathbf{K}} + \phi_{-\mathbf{H}-\mathbf{K}} \quad (2)$$

to be a triple,

$$Q_{\mathbf{LMN}} = \phi_{\mathbf{L}} + \phi_{\mathbf{M}} + \phi_{\mathbf{N}} + \phi_{-\mathbf{L}-\mathbf{M}-\mathbf{N}} \quad (3)$$

to be a quartet, and the functions $A_{\mathbf{HK}}$ and $B_{\mathbf{LMN}}$ to be:

$$A_{\mathbf{HK}} = \frac{2}{N^{1/2}}|E_{\mathbf{H}}E_{\mathbf{K}}E_{\mathbf{H}+\mathbf{K}}| \quad (4)$$

$$B_{\mathbf{LMN}} = \frac{2}{N}|E_{\mathbf{L}}E_{\mathbf{M}}E_{\mathbf{N}}E_{\mathbf{L}+\mathbf{M}+\mathbf{N}}|\,(|E_{\mathbf{L}+\mathbf{M}}|^2 + |E_{\mathbf{M}+\mathbf{N}}|^2 + |E_{\mathbf{N}+\mathbf{L}}|^2 - 2) \quad (5)$$

where N is the number of atoms, assumed identical in the whole unit cell, and I_1 and I_0 are modified Bessel functions. It should be noted that $B_{\mathbf{LMN}}$ can take on negative values when the cross terms ($|E_{\mathbf{L}+\mathbf{M}}|$, $|E_{\mathbf{M}+\mathbf{N}}|$, $|E_{\mathbf{N}+\mathbf{L}}|$) are very small, so it sums into the denominator of Eq. 1 as its absolute value. Further, the ratio of Bessel functions I_1/I_0 is known to be the expected value of the corresponding cosine. In view of Eqs. 2 and 3, Eq. 1 also defines R as a function, $R(\phi)$, of the phases. Because the magnitudes $|E|$ are presumed to be known, the functions $R(\Phi)$ and $R(\phi)$ are well defined solely as functions of Φ and ϕ, respectively.

The phases are functions, for a fixed choice of origin and enantiomorph, of the atomic position vectors. Specifically,

$$E_{\mathbf{H}} = |E_{\mathbf{H}}|e^{i\phi_{\mathbf{H}}} = \frac{1}{N^{1/2}}\sum_{j=1}^{N} e^{2\pi i \mathbf{H}\cdot\mathbf{r}_j} \quad (6)$$

where $\mathbf{r}_j$ is the position vector of the atom labeled j. Because the structure invariants $T_{\mathbf{HK}}$ and $Q_{\mathbf{LMN}}$ are uniquely determined for any given structure S, independent of the choice of origin, it follows that Eq. 1 also defines a function, $R(S)$, of structures S.

The Shake-and-Bake algorithm is targeted at minimizing the function in terms of the phases, while imposing the constraint of structural atomicity. As with any minimization strategy that is prone to locking in on local minima, our implementation will explore many initial structures (trials). Each initial structure is generated as a set of fractional atomic coordinates through random number generation. The generation of the random atomic coordinates is such that the resulting structures satisfy certain chemical constraints. Space group operators are then applied to the set of atoms in order to generate symmetry-related atomic positions. The resulting constellation of atoms is used in the structure factor calculation to arrive at a starting set of phases. The phase values are then adjusted by a local minimization procedure to reduce the value of $R(\Phi)$. After a minimization cycle, the adjusted phases are recombined with the measured structure factor amplitudes to calculate a Fourier map, through an inverse three-dimensional Fourier transform. This map is then scanned to locate the (at most) M highest peaks. These peaks constitute a new structure which has several favorable characteristics. It has (no more than) the requisite number of atoms, and it has been generated with the experimentally determined magnitudes. Currently, the resulting structure is recycled a fixed number of times through the process of Fourier transforma-

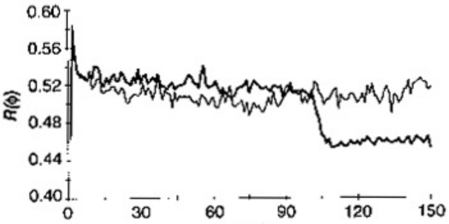

Fig. 3. The course of $R(\Phi)$ for a solution (bold) versus a nonsolution of ternatin(I) with respect to $R(\Phi)$.

Table 1. Data corresponding to solutions for both structures.

Parameter	Ternatin(I)	Ternatin(II)
Atoms in structure	110	105
Atoms per trial structure	104	104
Phases	1,000	1,000
Triples	20,000	20,000
Quartets	0	0
Reflections available	6,273	5,512
Reflections utilized	5,463	5,512
Cycles	150	150
Trials	2,048	2,048
Solutions	6	19
Percentage	0.3	0.9
Time per cycle(s)	22.5	28.5

tion, local minimization, Fourier synthesis, and peak picking. By observing the resulting values of $R(S)$ over the set of trials that have been processed, we are able to determine whether or not a solution has been obtained.

Based on experimentation with respect to the 28-atom, 84-atom, and 127-atom structures, we conjecture that the number of cycles of Shake-and-Bake necessary to determine the structures under consideration is of the order of 1.5 times the number of atoms in the structure. Therefore, we chose to perform the algorithm for 150 cycles on both of the previously unknown ~100-atom structures.

Experimentation on the 84-atom and 127-atom structures indicates that a cost-effective ratio for phases to atoms is approximately 10 to 1, while a cost-effective ratio for triples to phases is approximately 20 to 1, and the incorporation of negative quartets (that is, $B < 0$) may be unnecessary.

The experimentation described in this report has been performed predominantly on a Connection Machine CM-5 at Thinking Machines Corporation. Pertinent details of the experiments are given in Table 1. For both previously unknown structures, it was assumed that there were 104 atoms, although we subsequently found this not to be the case. Nevertheless, we used 104 atoms in the procedure. Further, based on the 10:1 phase to atom ratio and 20:1 triplet to phase ratio (no quartets), we chose to use 1,000 phases, 20,000 triples, and 0 quartets. Notice that in the case of ternatin(I), a number of reflections were removed from the full data set that corresponded to h indexes of 9 through 11 on the basis that their $F/\sigma(F)$ ratios were abnormally small. We chose to run the algorithm for 150 cycles using the 1.5:1 cycle to atom ratio. Based on available computer time, and desiring a sufficient sample size, we processed 2048 initial, randomly generated starting structures.

The six solutions produced for ternatin(I) had final $R(\Phi)$ values in the [0.45, 0.46] range, whereas the nonsolutions had final $R(\Phi)$ values greater than 0.49. The 19 solutions produced for ternatin(II) had final $R(\Phi)$ values in the [0.41, 0.42] range, whereas the nonsolutions had final $R(\Phi)$ values greater than 0.46. In other words, as mentioned previously, $R(\Phi)$ is diagnostic in terms of detecting solutions. A visual representation of the convergence of a solution versus a nonsolution for ternatin(I) with respect to $R(\Phi)$ is shown in Fig. 3. In fact, based solely on the final $R(\Phi)$ values, we were able to determine that after 64 trials of ternatin(I) a single solution was at hand, and that after 64 trials of ternatin(II) there were two solutions. Each initial 64-trial experiment was performed in ~70 CPU min on a 64-node Connection Machine CM-5. It was only later that we decided to run both structures for 2048 trials for statistical purposes. The percentage of success was significantly higher for ternatin(II) than for ternatin(I). This difference may be due to the fact that there was a threefold higher percentage of aberrant triples with high A values for ternatin(I) as compared to ternatin(II), which more nearly matched the expected rate of failure predicted by the A-values.

REFERENCES AND NOTES

1. G. DeTitta *et al.*, in *Proceedings of the Sixth Distributed Memory Computing Conference* (IEEE Computer Society, New York, 1991), pp. 587–594.
2. C. M. Weeks, G. T. DeTitta, R. Miller, H. A. Hauptman, *Acta Crystallogr. D* **49**, 179 (1993).
3. N. Bashir *et al.*, in *Proceedings of the Fifth Distributed Memory Computing Conference* (IEEE Computer Society, New York, 1990), pp. 513–521.
4. H. A. Hauptman, *Abstr. Am. Crystallogr. Assoc.* **16**, 53 (R4) (1988).
5. H. A. Hauptman, *Crystallographic Computing 5: From Chemistry to Biology*, D. Moras, A. D. Podjarny, J. C. Thierry, Eds. (Oxford Univ. Press, New York, 1991), pp. 324–332.
6. C. Weeks, W. Duax, M. Wolff, *Acta Crystallogr. B* **32**, 261 (1976).
7. V. Z. Pletnev, N. Galitsky, G. Smith, C. Weeks, W. Duax, *Biopolymers* 19, 1517 (1980).
8. V. Z. Pletnev, V. T. Ivanov, D. A. Langs, P. Strong, W. L. Duax, *ibid.* **32**, 819 (1992).
9. Crystals were grown by slow evaporation from dioxane. Data were recorded in 1982 at room temperature with CuKα radiation on a Syntex P1 diffractometer to a resolution of 0.94 Å in which a 96-step $\theta - 2\theta$ scanning procedure was used. Crystal data: $2(C_{37}H_{67}N_7O_8)\cdot C_4H_8O_2$ orthorhombic $P2_12_12_1$, a = 11.563(1), b = 21.863(2), c = 36.330(4) Å (numbers in parentheses are errors in the last digit), and $Z = 4$.
10. Crystals were grown by slow evaporation from 95% ethanol. Data were recorded in 1990 at 153 K with CuKα radiation on a Nonius CAD4 diffractometer to a resolution of 0.97 Å in which a similar $\theta - 2\theta$ step scan procedure was used. Crystal data: $2(C_{37}H_{67}N_7O_8)\cdot H_2O$, orthorhombic $P2_12_12_1$, a = 14.067(2), b = 16.695(1), c = 36.824(4) Å, and $Z = 4$.
11. J.-X. Yao, *Acta Crystallogr. A* **37**, 642 (1981).
12. M. G. Rossmann, Ed., *The Molecular Replacement Method* (Gordon and Breach, New York, 1972).
13. R. Miller *et al.*, unpublished results.
14. T. Debaerdemaeker, C. Tate, M. M. Woolfson, *Acta Crystallogr. A* **44**, 353 (1988).
15. G. M. Sheldrick, *ibid.* **46**, 467 (1990).
16. M. Shiono and M. M. Woolfson, *ibid.* **48**, 451 (1992).
17. E. H. L. Aarts and P. J. M. van Laarhoven, *Simulated Annealing: Theory and Applications* (Reidel, New York, 1988).
18. D. E. Holland, *Genetic Algorithms in Search, Optimization, and Machine Learning* (Addison-Wesley, Reading, MA, 1989).
19. C.-S. Chang *et al.*, *Int. J. Supercomput. Appl.*, in press.
20. Supported by NIH grants GM-46733, GM-32812, and DK-19856 (in part), NSF grant IRI-9108288 and the Harker grant. We thank N. Galitsky for x-ray diffraction data collection; C.-S. Chang, A. Khalak, C.-W. Lee, and P. Thurman, who performed some of the computational experiments described in this report; and the Thinking Machines Corporation for allowing us to use their CM-5 machines.

2 October 1992; accepted 5 January 1993